Anal. Chem. **2000,** *72,* 101−109

QSPR Correlation and Predictions of GC Retention Indexes for Methyl-Branched Hydrocarbons Produced by Insects

Alan R. Katritzky* and Ke Chen

Center of Heterocyclic Compounds, Department of Chemistry, University of Florida, P.O. Box 117200, Gainesville, Florida 32611-7200

Uko Maran

Department of Chemistry, University of Tartu, 2 Jakobi St., Tartu 51014, Estonia

David A. Carlson

Center for Medical, Agricultural and Veterinary Entomology, United States Department of Agriculture, Agricultural Research Service, P.O. Box 14565, Gainesville, Florida 32604

A successful interpretation of the complex manner by which the GC retention indexes of methylalkanes produced by insects are related to chemical structure was achieved using the quantitative structure−property relationship (QSPR) method. A general QSPR model including mainly topological descriptors was obtained for 178 data points. The error of the model is similar to the experimental error. The model was supported by (i) leave-one-out cross validation and (ii) division into three sets and prediction of each set from the other two. As a further test of the utility of the model, retention indexes were successfully predicted for an external set of 30 methyl-branched hydrocarbons not involved in the deduction of the correction equation from the main data set. General trends of the structural variation of compounds in any given range of retention index are discussed. The average error was 4.6 overall and 4.3 for the 165 compounds remaining after leaving out small monomethyl alkanes.

Insects produce a great variety of methyl-branched alkanes,[1] but the structural differences between them are generally quite limited. Typically, most of the alkanes and methylalkanes synthesized by insects have a straight-chain backbone of 21−37 carbons, but this may extend to 51 carbons. A majority of the methylalkanes have odd-numbered carbon chains for backbones. The methyl branches in those backbones appear at restricted locations. Most insects produce monomethyl alkanes with the methyl branch located on carbon 2, 3, 7, 9, 11, 13, or 15. The next most commonly found series consists of dimethyl alkanes, in which the methyl branches are most often separated by a chain of 3 methylene (−CH_2−) groups and can also be separated by 7, 9, or 11 −CH_2− groups; in these dimethyl derivatives the methyl branches are almost never separated by an even-number of carbons. The same pattern appears for the trimethyl alkanes, where three methyl branches separated by chains of three −CH_2− groups are the most common. In tetramethyl alkanes, those with the four methyl branches separated by three −CH_2− groups are the only types observed so far.

The alkanes and methylalkanes described above are usually considered to be waterproofing agents present on the cuticle, the hard chitinous body covering or exoskeleton. These body-surface components may also contain specific attractive chemical compounds that cause aggregation and/or sex pheromone activity. It is important to determine their chemical structures to make more effective lures. The principal method used for the identification of these alkanes is gas chromatography (GC) and GC-mass spectrometry (GC-MS).[1,2] However, the interpretation of their GC-MS spectra is problematic, because of (i) the difficulties in interpreting similar or overlapping GC peaks, (ii) similarity between the MS of similar methyl-branched alkanes, and (iii) the fact that the mass spectra of few of these compounds have been properly entered into useful databases. In addition, the algorithms commonly used for compound selection from the databases do not select between candidates with useful accuracy. Nevertheless, the previous work by one of our groups[1] indicated that the various classes of methylalkanes seemed to have internally consistent retention times for analogues and homologues of similar molecular structures on apolar columns. Therefore, it seemed worthwhile to attempt a correlation of the methylalkane structures with their retention times or retention indexes in order to correlate and, hopefully, to be able to predict quantitatively such properties for this type of methylalkane by structural molecular descriptors.

Quantitative structure−property relationships (QSPR) have been demonstrated to be a powerful tool in chromatography.

* Corresponding author: (tel.) 352-392-0554; (fax) 352-392-9199; (e-mail) katritzky@chem.ufl.edu.

(1) Carlson, D. A.; Bernier, U. R.; Sutton, B. D. *J. Chem. Ecol.* **1998**, *24*, 1845−1865.

(2) Kissin, Y. V.; Feulmer, G. P. *J. Chromatogr. Sci.* **1986**, *24*, 53−59.

10.1021/ac990800w CCC: $19.00
Published on Web 12/01/1999

Table 1. Reference Overview of QSPR Models on GC Retention Index

entry	class of compounds	R^2	s	n	N_D	classes of descriptors	ref
1	alkylbenzenes	0.996	6.54	41	5	boiling point, topological, and quantum chemical	3
2	sulfur vesicants	0.996	31.5	31	4	topological, constitutional, and geometrical	4
3	hydrocarbons	0.966	18.6	67	4	topological, constitutional, and geometrical	5
4	isoalkanes	0.999	1.6	38	4	physicochemical, topological, and geometrical	6
5	polycyclic aromatic hydrocarbons	0.947	0.159	31	3	the moment of inertia, solute length-to-breadth ratio, and connectivity index	7
6	flavonoids	0.951	0.120	49	5	topological, geometric, and electronic	8
7	alkylbenzenes	0.968	24.5	150	6	topological, electronic, and geometrical	9
8	polycyclic aromatic hydrocarbons	0.994	7.10	100	2	quasilength of carbon chain and pseudo-conjugated system surface	10
9	organic compounds	0.981	19.2	381	16	topological	11
10	organic compounds	0.959	0.515	152	6	quantum chemical, constitutional, electronic	19
11	polyalcylated pyridines	0.971	17.8	50	6	constitutional, topological, quantum chemical	21
12	methylalkanes	0.959	5.8	178	4	topological descriptors	**present work**

QSPR have been used to obtain simple models to explain and predict the chromatographic behavior of various classes of compounds. Recent work in this area is summarized and some of the models are listed in Table 1. Mekenyan et al. derived linear quantitative retention-structure models in gas chromatography for 41 alkylbenzenes: boiling point, two geometric indexes, and two electronic structure descriptors in their best equation (Table 1: no. 1).[3] Jurs and his group correlated the observed Kovats retention indexes of sulfur vesicants by multiple linear regression techniques using 9 descriptors (topological, electronic, and geometrical) in the models for different stationary phases (Table 1: no. 2, only one example is given).[4] Later, they predicted the GC retention behavior of 67 hydrocarbons. Several models were developed for two stationary phases using interactive regression analysis. The geometrical and topological descriptors gave good results with 4 descriptors (Table 1: no. 3). It was also found that the boiling point is a successful physicochemical descriptor when combined with structure-based descriptors.[5]

Dimov and Osman used a quantitative structure-retention relationship to relate the chromatographic retention of 38 isoalkanes to their molecular structural features. The descriptor contributions were divided into two groups: basic and tuning which allows better orientation in retention modeling and better understanding of the retention connected with molecule structure. They found that quantum chemical calculations of conformation states increase the predictive accuracy in absorption mode chromatography (Table 1: no. 4).[6] Welsh et al. studied the chromatographic data from 31 unsubstituted 3−6 ring polycyclic aromatic hydrocarbons using CoMFA in quantitative structure-retention relationship studies. They took the moment of inertia as a basis for CoMFA alignment of their data set (Table 1: no. 5).[7]

Payares and co-workers studied gas chromatographic behavior of 49 flavanoids in an apolar column. They found that the topological descriptor $(1/({}^3\chi_c - {}^3\chi_c v))$ and the sum of the values of the charges for the hydroxyl hydrogens are the best descriptors in the QSRR model (Table 1: no. 6).[8] Jurs and co-workers[9] developed a QSPR model of retention indexes based on the structure of 150 alkylbenzenes using topological, geometric, and electronic descriptors (Table 1: no. 7). Kang et al. successfully predicted the capillary GC retention indexes of 100 polycyclic aromatic hydrocarbons by using two parameters: pseudo-conjugated π-system surface and quasilength of carbon chain with a correlation coefficient of 0.9948 (Table 1: no. 8).[10]

Pompe and Novič used an extensive data set of 381 simple organic compounds and topological descriptors calculated with the CODESSA program to predict retention indexes. They also compared three different methods: multiple linear regression (MLR) approach, back-propagation (BP), and counterpropagation (CP) artificial neural networks (NN). The MLR model (Table 1: no. 9) with 16 descriptors in the model outperformed CP NN but was slightly worse than the BP NN approach.[11] Chrétien and co-workers[12] applied factor analysis to retention (Kovats) indexes for three congeneric aromatic series of substituted benzene, benzaldehyde, and acetophenone compounds. Kupchik[13,14] obtained structure−gas chromatographic retention time models for 26 tetra-*n*-alkylgermanes, 26 tetra-*n*-alkylsilanes, and a mixed set of silanes and germanes using topological indexes, which include molecular connectivity indexes, the Kier-Hall total topological indexes, and the Kier-Hall electrotopological state atom index for the silicon and germanium atoms.

In recent years, methodology for a general QSPR approach has been developed and coded as the CODESSA software package which combines different ways of quantifying the structural information about the molecule with advanced statistical analyses for the establishment of molecular structure−property relationships. CODESSA can calculate a large number of quantitative descriptors solely on the basis of molecular structural information.[15,16,17] CODESSA has been applied successfully to predict a variety of physical properties of compounds.[18]

(3) Mekenyan, O.; Dimov, N.; Enchev, V. *Anal. Chim. Acta* **1992**, *260*, 69−74.
(4) Woloszyn, T. F.; Jurs, P. C. *Anal. Chem.* **1992**, *64*, 3059−3063.
(5) Woloszyn, T. F.; Jurs, P. C. *Anal. Chem.* **1993**, *65*, 582−587.
(6) Dimov, N.; Osman, A. *Anal. Chim. Acta* **1996**, *323*, 15−25.
(7) Collantes, E. R.; Tong, W.; Welsh, W. J.; Zielinski, W. L. *Anal. Chem.* **1996**, *68*, 2038−2043.
(8) Payares, P.; Diaz, D.; Olivero, J.; Vivas, R.; Gomez, I. *J. Chromatogr., A* **1997**, *771*, 213-219.
(9) Sutter, J. M.; Peterson, T. A.; Jurs, P. C. *Anal. Chim. Acta* **1997**, *342*, 113−122.
(10) Kang, J.; Cao, C.; Li, Z. *J. Chromatogr. A* **1999**, *799*, 361−367.
(11) Pompe, M.; Novič, M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 59−67.
(12) Righezza, M.; Hassani, A.; Meklati, B. Y.; Chrétien, J. R. *J. Chromatogr.* **1996**, *723*, 77−91.
(13) Kupchik, E. J. *J. Chromatogr.* **1993**, *630*, 223−230.
(14) Kupchik, E. J. *J. Chromatogr.* **1993**, *645*, 182−184.
(15) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *CODESSA Version 2.0 Reference Manual*, University of Florida: Gainesville, Florida, 1994.

Our groups have already utilized CODESSA for QSPR of gas-chromatographic properties. A general QSPR treatment on 152 individual structures incorporating a wide cross section of classes of organic compounds provided good six-parameter correlations for gas-chromatographic retention times (Table 1: no. 10, $R^2 = 0.959$, $s = 0.515$ for t_R) and for Dietz flame-ionization response factors ($R^2 = 0.892$, $s = 0.0543$ for RF_{Dietz}).[19] In the case of t_R, the most important descriptors were α-polarizability and the minimum valency at any H atom, describing the dispersional and hydrogen-bonding interaction between the compound studied and the gas-chromatographic solid medium, respectively. In the case of RF, the most important descriptors were found to be the relative weight of the "effective" carbon atoms and the total molecular one-center one-electron repulsion energy in the molecule. The possibility of predicting values is of particular significance for the response factors, which are independent of GC column parameters. A new efficient approach for variable selection based on multiregression has been used to predict t_R (retention times) and RF (response factor) for the same data set.[20] Another successful six-parameter QSPR model was obtained for the retention indexes of 50 polyalkylated pyridines (Table 1: no. 11, $R^2 = 0.971$, $s = 17.8$).[21,22] The descriptors involved in the equation reflect the relative position and size of alkyl groups connected to the pyridine ring. They also show the importance of intermolecular interactions between solute and stationary phase, upon which gas-chromatographic retention depends.

The primary purpose of the present work is to establish QSPR models of retention indexes for a large set of methylalkanes produced by insects. It is expected that the retention indexes of methylalkanes could be predicted using definite molecular structural descriptors. The descriptors chosen in the models should reflect the relative positions and the number of the multiple methyl groups attached to the carbon chain, the conformation of the compound, and the length of the carbon backbone. If so, the model should help select between possible structures for any specific value of the retention index. We now report our results utilizing topological descriptors and quantum chemical descriptors and explain how they justify these expectations.

EXPERIMENTAL SECTION

Data Set. The data sets of the Kovats retention indexes were chosen from refs 1 and 2. A total of 178 methylalkanes (Table 2), including monomethylalkanes, dimethylalkanes, trimethylalkanes, and tetramethylalkanes, comprised a range of different carbon chain lengths. Retention indexes in the data set fall in the range 27.3−74.5 for monomethyl alkanes; 55−110 for di-methyl alkanes; 71−138 for tri-methyl alkanes; and 81−162 for tetra-methyl alkanes. Most of the compounds are natural products, but a few were synthesized for other purposes. Additionally, an external data set of 30 compounds (Table 3) was measured to test the predictive quality of the QSPR model. The retention indexes of all compounds were determined by GC and GC-MS under a single set of conditions, such that all parameters were as internally consistent as possible. An apolar fused silica capillary column (DB-1 30 m × 0.32 mm i.d., 0.25 μm phase thickness), on-column injection, and temperature programming from 60−320 °C were used to ensure good chromatography. All data points were obtained from extended temperature programmed measurements of natural compounds, for which there are very few synthetic equivalents. *n*-Alkane standards are not available for all carbon numbers for the larger compounds. Different quantities of methyl branched alkanes were present in these natural samples, presenting a problem with overloading of some peaks and delayed elution; the published experimental values used here were not adjusted. These factors could lead to some small but systematic experimental errors for the larger compounds cited. Only the last two digits of the KI were recorded in Table 2. These two-digit values were obtained after subtracting the number of the carbons in the main chain × 100, e.g., from KI (3133 − 3100 = 33) or (3855 − 3700 = 155). Co-injected *n*-alkanes were employed to determine these values within ± 2 units.

Computational Methods. All the molecular structures were drawn and preoptimized by the MMX molecular mechanics method incorporated into the PCMODEL program.[23] The final structural optimizations of compounds were performed using the AM1 parametrization[24] within the semiempirical quantum-chemical program MOPAC 6.0.[25] Thereafter CODESSA program was used to calculate five types of molecular descriptors: constitutional, topological, geometrical, electrostatic and quantum-chemical.[16] Altogether, 302 descriptors were calculated for each of the 178 compounds studied. The correlation analysis to find the best QSPR model was carried out using the heuristic method in the CODESSA program. This procedure is based on the scale forward selection technique[26] and has been described in detail elsewhere.[21,27] However, the branching criterion was changed to 10 instead of the default, three. This extended the spectrum of best-two parameter correlations from which the correlation equations with higher numbers of descriptors were developed.

The structure of the data set is simple, consisting solely of methyl-branched alkanes, which have only carbon and hydrogen atoms in the chain with neither functional groups nor heteroatoms. The main differences between the compounds are the length of the chain, the positions of the methyl groups, and the number of the methyl groups connected to the backbone. The columns used for GC are nonpolar, and as the compounds do not contain any heteroatoms, the dipole−dipole interactions are not expected to be important in solute−stationary phase and solute−solute interactions. Preliminary analysis of the descriptors and development of QSPR models reveals that descriptors related to charge and dipole moments, and consequently, polar effects between

(16) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *Chem. Soc. Rev.* **1995**, 279−287.
(17) Katritzky, A. R.; Karelson, M.; Lobanov, V. S. *Pure Appl. Chem.* **1997**, *69*, 245−248.
(18) Karelson, M.; Maran, U.; Wang, Y.; Katritzky, A. R. *Collect. Czech. Chem. Commun.* **1999**, *64*, 1551−1571.
(19) Katritzky, A. R.; Ignatchenko, E. S.; Barcock, A. R.; Lobanov, V. S.; Karelson, M. *Anal. Chem.* **1994**, *66*, 1799−1807.
(20) Lucic, B.; Trinajstic, N.; Sild, S.; Karelson, M.; Katritzky, A. R. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 610−621.
(21) Katritzky, A. R.; Lobanov, V.; Karelson, M.; Murugan, R.; Grendze, M. P.; Toomey, J. E. *Rev. Roum. Chim.* **1996**, *41*, 851−867.
(22) Murugan, R.; Grendze, M. P.; Toomey, J. E.; Katritzky, A. R.; Karelson, M.; Lobanov, V.; Rachwal, P. *CHEMTECH* **1994**, *24*, 17−23.
(23) *PCMODEL User Manual*; Serena Software: Bloomington, IN, 1992.
(24) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(25) Stewart, J. J. P. "*MOPAC 6.0*"; QCPE No 455, 1990.
(26) Draper, N. R.; Smith, H. *Applied Regression Analysis*; Wiley: New York, 1966.
(27) Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. *J. Phys. Chem.* **1996**, *100*, 10400−10407.

Table 2. Comparison of the Experimental Retention Indexes for 178 Mono, Di-, Tri-, and Tetramethylalkanes Produced by Insects, with their Prediction by the QSPR Model

structure[a]	retention index exp	retention index calc	Δ	structure[a]	retention index exp	retention index calc	Δ	structure[a]	retention index exp	retention index calc	Δ
					Monomethyl						
09_02	66.5	65.9	−0.6	23_02	64.0	71.0	7.0	31_06	43.2	41.2	−2.0
09_03	73.0	53.4	−19.6	23_03	74.5	80.3	5.8	31_07	40.0	38.2	−1.8
11_02	66.5	74.1	7.6	23_12	37.0	31.5	−5.5	31_13	30.8	29.8	−1.0
11_03	72.5	66.2	−6.3	25_02	63.0	68.1	5.1	31_16	29.8	28.7	−1.1
13_02	66.5	77.5	11.0	25_03	74.4	79.6	5.2	33_02	62.0	58.5	−3.5
13_03	73.0	74.1	1.1	25_13	34.5	30.9	−3.6	33_03	74.5	73.8	−0.7
15_02	66.5	78.3	11.8	27_02	63.0	65.9	2.9	33_04	57.5	51.1	−6.4
15_03	73.7	79.0	5.3	27_03	74.4	78.3	3.9	33_05	50.0	42.4	−7.7
17_02	65.8	77.4	11.6	27_14	33.0	30.3	−2.7	33_06	43.7	40.3	−3.4
17_03	74.0	81.1	7.1	29_02	62.2	63.4	1.2	33_13	28.5	29.5	1.0
19_02	66.0	75.5	9.5	29_03	74.0	76.8	2.8	33_17	28.5	27.9	−0.7
19_03	74.3	81.8	7.5	29_15	31.5	29.5	−2.0	35_02	62.0	56.2	−5.8
19_10	43.0	31.6	−11.4	31_02	61.5	60.8	−0.7	35_03	74.3	72.1	−2.3
21_02	66.0	73.4	7.4	31_03	74.1	75.3	1.2	35_18	27.3	27.0	−0.3
21_03	74.5	81.5	7.0	31_04	57.5	52.3	−5.2				
21_11	41.0	31.7	−9.3	31_05	50.0	43.4	−6.6				
					Dimethyl						
22_0822	67.0	69.0	2.0	29_0313	104.0	107.2	3.2	33_0517	80.0	75.0	−5.0
23_0309	110.0	111.6	1.6	29_0513	82.0	78.0	−4.0	33_0519	82.0	75.6	−6.4
24_0509	85.0	82.0	−3.0	29_0519	83.0	80.1	−2.9	33_0717	70.0	69.8	−0.2
25_0311	109.0	110.0	1.0	29_0717	73.0	72.3	−0.7	33_1123	62.4	67.7	5.3
25_0315	105.0	110.7	5.7	30_0206	105.0	97.8	−7.2	34_0210	94.0	88.9	−5.1
25_0511	82.0	79.9	−2.1	30_0210	99.0	92.2	−6.8	34_0416	89.0	82.8	−6.2
25_0517	85.0	83.5	−1.5	30_0212	95.0	90.4	−4.7	34_0610	73.8	76.6	2.8
25_0711	77.0	72.9	−4.2	30_0307	108.0	113.2	5.2	34_0812	65.0	70.0	5.0
26_0206	104.0	101.4	−2.6	30_0410	94.0	89.2	−4.8	34_1222	61.4	64.3	2.9
26_0408	95.0	93.8	−1.2	30_0610	75.0	77.9	2.9	34_1317	55.0	61.0	6.0
26_0511	82.0	79.9	−2.1	31_0307	109.0	113.7	4.7	35_0307	109.5	110.5	1.0
26_0610	78.0	78.0	0.0	31_0313	103.5	106.1	2.6	35_0315	101.0	102.6	1.6
26_0711	75.0	73.1	−1.9	31_0315	109.0	105.0	−4.0	35_0509	80.0	79.0	−1.0
27_0307	109.0	114.7	5.7	31_0513	80.5	77.2	−3.3	35_0519	80.5	73.7	−6.8
27_0315	105.0	108.6	3.6	31_0517	82.0	76.6	−5.5	35_0717	69.7	68.7	−1.0
27_0511	82.0	79.8	−2.2	31_0711	70.2	73.0	2.8	35_0921	61.0	66.5	5.5
27_0517	86.0	80.7	−5.3	31_1121	62.9	65.6	2.7	36_0212	95.0	85.3	−9.7
27_0713	74.0	71.9	−2.1	32_0208	97.0	93.0	−4.0	36_0517	80.0	72.9	−7.1
27_0919	65.0	72.6	7.6	32_0408	92.0	91.2	−0.8	36_1323	61.0	62.4	1.4
28_0206	105.0	99.7	−5.3	32_0610	73.5	77.4	3.9	37_0315	101.0	101.3	0.3
28_0210	99.0	93.7	−5.3	32_0812	66.0	70.4	4.4	37_0509	79.0	77.8	−1.2
28_0410	95.0	89.9	−5.1	32_0921	62.0	69.4	7.4	37_0517	80.0	72.2	−7.8
28_0515	82.0	78.2	−3.8	32_1418	57.5	61.2	3.7	37_1323	59.0	61.5	2.5
28_0713	73.0	71.8	−1.2	33_0309	103.0	109.3	6.3	38_0517	78.0	71.6	−6.4
29_0307	108.0	114.0	6.0	33_0315	109.0	103.6	−5.4				
					Trimethyl						
24_040812	120.0	118.3	−1.7	32_061418	99.0	98.9	−0.1	35_131721	77.0	85.9	8.9
25_050913	110.0	106.4	−3.6	32_121620	81.0	88.1	7.1	35_131723	83.0	85.9	2.9
26_040812	119.0	118.2	−0.8	33_030715	136.5	134.8	−1.7	36_040816	115.0	113.4	−1.6
27_030711	138.0	138.3	0.3	33_051317	105.0	100.8	−4.2	36_081216	85.0	94.0	9.0
28_040812	118.0	118.1	0.1	33_071115	89.0	97.1	8.1	36_141822	76.0	84.8	8.8
29_030711	137.0	138.0	1.0	33_111519	79.6	88.4	8.8	37_030715	135.0	132.4	−2.6
29_051317	107.0	103.3	−3.7	34_021016	124.0	111.4	−12.6	37_051317	103.0	98.9	−4.1
30_061418	100.0	100.0	−0.1	34_040812	115.5	116.3	0.8	37_071319	84.0	94.1	10.1
31_030711	136.5	137.6	1.1	34_061418	97.0	97.7	0.7	37_151923	75.0	84.0	9.0
31_051317	105.4	101.9	−3.5	34_081216	86.4	94.4	8.0	38_162024	73.5	83.2	9.7
31_071317	91.3	96.3	5.0	34_121620	78.0	87.2	9.2	39_051317	101.0	97.9	−3.1
31_111519	81.0	89.4	8.4	35_030715	136.3	133.6	−2.7	39_151923	72.4	82.5	10.1
32_021016	124.0	112.9	−11.1	35_050913	105.0	105.0	0.0	40_141822	71.0	82.6	11.6
32_041216	116.0	110.6	−5.4	35_071115	88.3	97.0	8.7				
					Tetramethyl						
29_03071115	162.0	156.4	−5.6	35_07111519	128.0	116.4	−11.6	37_03071115	155.0	154.5	−0.6
31_03071115	161.0	156.5	−4.5	35_09131721	117.0	111.7	−5.3	37_07111519	123.0	115.9	−7.2
31_04081216	149.0	136.7	−12.3	35_11151924	105.0	109.2	4.2	37_09131721	113.0	110.6	−2.4
33_03071115	159.0	155.8	−3.2	36_06101216	123.0	118.2	−4.8	37_11151924	103.0	107.7	4.7
33_04081216	148.0	136.3	−11.7	36_08121620	113.0	113.7	0.7	38_10141822	100.0	108.7	8.7
35_03071115	158.0	155.3	−2.7	36_10141822	103.5	109.8	6.3				

[a] The first two digits show the length of the carbon backbone, each next two digits show the position of a methyl substituent on the backbone.

Table 3. Prediction of Retention Index for External Test Set of 30 Methyl-Branched Alkanes Utilizing the QSPR Model in Table 4

	retention index		
structure	exp	predicted	Δ
27_05	50.3	45.5	−4.8
29_07	39.8	38.8	−1.0
21_0711	72.0	71.2	−0.8
23_0311	105.0	110.0	−5.0
25_0307	108.5	115.6	7.1
25_0509	86.0	82.2	−3.8
26_0410	92.5	90.4	−2.1
26_0613	81.0	75.8	−5.2
27_0515	83.2	78.9	−4.3
27_0711	67.2	73.3	6.1
27_0911	65.0	67.8	2.8
28_0408	95.0	93.2	−1.8
29_0509	82.0	81.8	−0.2
31_0719	66.0	72.2	6.2
31_0919	65.0	68.5	3.5
32_0210	91.0	90.6	−0.4
34_0212	94.0	87.0	−7.0
34_0614	75.0	73.0	−2.0
27_030713	140.0	137.7	−2.3
28_021018	118.0	118.3	0.3
29_091317	95.0	92.9	−2.1
31_050913	100.0	106.2	6.2
31_071115	91.3	97.5	6.2
31_091317	92.2	92.2	0.0
33_050923	109.0	109.4	0.4
33_071317	95.0	95.5	0.5
33_091317	91.9	91.8	−0.1
34_061014	96.0	103.1	7.1
34_061216	100.0	101.8	1.8
34_101418	89.9	89.8	−0.1

stationary phase and mobile phase, do not contribute essential information from the point of the interaction mechanism for the present set of compounds. The presence of such descriptors related to charge and dipolar moment increases the possibility of chance correlations and may exclude descriptors that could be important. Therefore, we reduced our set of descriptors to 129, by excluding 70 electrostatic descriptors (related to charge distribution and hydrogen bonding in the molecule) and 103 of the quantum chemical descriptors (HOMO and LUMO energy-related descriptors and charge, dipole, and hydrogen-bonding related descriptors). The remaining 129 descriptors differentiate the structures and represent the molecule structure in such a way to enable us to obtain a prediction model that is as simple as possible.

RESULTS AND DISCUSSION

The best four-parameter correlation equation obtained for the whole data set of 178 compounds is presented in detail in Table 4 and Figure 1 with squared correlation coefficient $R^2 = 0.9585$. The standard deviation of the model is $s = 5.8$, and the average error is 4.6, close to the average experimental error of 4. The cross-validated correlation coefficient $R_{cv}^2 = 0.9543$, in comparison with correlation coefficient R^2, indicates the stability of the QSPR model. To further demonstrate the absence of chance correlation, the whole data set was divided into three subsets (by using numbers 1, 4, 7, etc; 2, 5, 8, etc; and 3, 6, 9, etc) and each subset was predicted by using the other two subsets as the training set. In this procedure, the same descriptors were retained in the correlation equation, but the coefficients were allowed to vary. The results are shown in Table 4, with average training quality of $R^2 = 0.9588$ and average predicting quality of $R^2 = 0.9553$, which indicates that proposed model has a high statistical stability and validity. The average predicting quality is even higher ($R^2 = 0.9603$) when the biggest outlier, 3-methylnonane, is removed from predicted set no. 2 (Table 5).

$$^{k}\mathrm{AIC} = -\sum_{n}^{n_i} \log_2 \frac{n_i}{n} \qquad (1)$$

The first descriptor in the model is the Average Information Content (1st order) denoted as ^{1}AIC, which also has the highest single parameter correlation $R^2 = 0.4627$. The ^{1}AIC is defined on the basis of the Shannon information theory and is calculated as given in eq 1,[28] where n_i is the number of atoms in the ith class and n is the total number of atoms in the molecule. The atoms are divided into different classes by taking into account the coordination sphere. This leads to the information content indexes of different order k. In the current case, for ^{1}AIC, the class is one, meaning that the coordination sphere covers only the first valence level, considering atoms with similar first-order neighbors, directly connected to this atom, in the molecular graph.[29] In essence, this descriptor gives us information on how many atoms with a similar connectivity pattern we have in the molecule. The current set of compounds contains four different connectivity patterns in the molecule, namely C−HHHC, C−HHCC, C−HCCC, and H−C. The descriptor is dependent on the number of atoms involved in the molecule, and it arranges the molecules in the order of rising chain length and number of the substituents of aliphatic alkanes.

$$^{k}\mathrm{ASIC} = \frac{^{k}\mathrm{IC}}{\log_2 n} \qquad (2)$$

The second descriptor is Average Structural Information Content (2nd order), denoted as, ^{2}ASIC. Its definition is based on the same principles as the previous descriptor, and it can be calculated according to eq 2,[30] where the coordination sphere, k, is two and counts up to second-order neighbors on the molecular graph. In addition to the first valence level, the second valence level can differentiate up to 7 different connectivity patterns in the molecule, depending on the position of the substituents, namely: C*−HHHC(HCC), C*−HC(HHH)C(HHH)C(HHC), C*−HHC(HCC)C(HHC), C*−HHC(HHC)C(HHC), H*−C(HHC), H*−C(CCC), and H*−C(HCC) (See Figure 6). The descriptor may be looked at as a normalized information content, with the maximum information content, $\log_2 n$, as the normalization factor.[31]

(28) Kier, L. B. *J. Pharm. Sci.* **1980**, *69*, 807−810.
(29) Basak, S. C.; Niemi, G. J.; Veith, G. D. *J. Math. Chem.* **1990**, *4*, 185−205.
(30) Basak, S. C.; Harriss, D. K.; Magnuson, V. R. *J. Pharm. Sci.* **1984**, *73*, 429−437.
(31) Roy, A. B.; Basak, S. C.; Harriss, D. K.; Magnuson, V. R. In *Mathematical, Modelling in Science and Technology*; Avula, X. J. R., Kalman, R. E., Liapis, A. I., Rodin, E. Y., Eds.; Pergamon Press: New York, 1983; pp 745−750.

Table 4. The QSPR Model for Retention Indexes of 178 Methylalkanes Produced by Insects[a]

no.	coefficient	error	R^2	R^2_{cv}	s	F-test	t-test	descriptor
0	$-6.5969 \times 10^{+04}$	$2.7530 \times 10^{+03}$					-23.9624	intercept
1	$1.3834 \times 10^{+03}$	$3.6128 \times 10^{+01}$	0.4627	0.4472	20.62	151.54	38.2926	^{1}AIC
2	$-1.2040 \times 10^{+03}$	$3.7078 \times 10^{+01}$	0.6520	0.6413	16.64	163.94	-32.4722	^{2}ASIC
3	$1.8914 \times 10^{+03}$	$7.9919 \times 10^{+01}$	0.8794	0.8746	9.83	422.83	23.6670	$E^{\max}_{\text{el.}-\text{el.repuls.}}(C-H)$
4	$-1.3067 \times 10^{+02}$	$7.1912 \times 10^{+00}$	0.9585	0.9543	5.78	999.62	-18.1710	Balaban index

[a] The intercorrelations of the four descriptors are available in the supporting material.

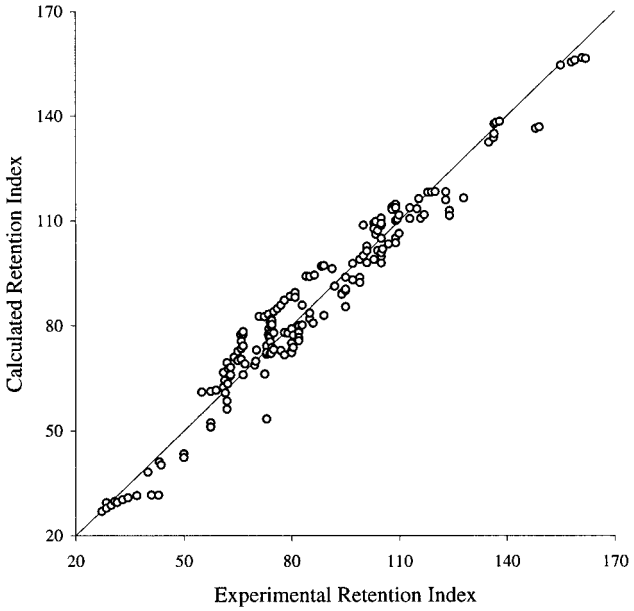

Figure 1. Plot of the predicted values for 178 methylalkanes vs the experimental retention indexes.

Table 5. Verification of Statistical Validity of the Model

training sets	R^2	predicted sets	R^2
2 & 3	0.9549	1	0.9645
1 & 3	0.9649	2	0.9394 (0.9544)
1 & 2	0.9565	3	0.9619
average	0.9588	average	0.9553 (0.9603)

The descriptor shows how branched the molecule is and how complex the neighborhood of the various carbon atoms is.

$$E^{\max}_{\text{el.}-\text{el.repuls.}}(A-B) = \sum_{\mu,\nu\in A}\sum_{\lambda,\sigma\in B} P_{\mu\nu}P_{\lambda\sigma}\langle\mu\nu|\lambda\sigma\rangle \qquad (3)$$

The third descriptor, the maximum electron−electron repulsion on C−H bond ($E^{\max}_{\text{el.}-\text{el.repuls.}}(C-H)$), belongs to the class of quantum chemical energy related descriptors. It describes the energy distribution in the molecule and is calculated as defined in eq 3, where $P_{\mu\nu}$ and $P_{\lambda\sigma}$ are the density matrix elements and $\langle\mu\nu|\lambda\sigma\rangle$ are the electron-repulsion integrals on the atomic basis $\{\mu\nu\lambda\sigma\}$. The electron−electron repulsion energy describes the electron-repulsion-driven processes in the molecule and may be related to the conformational (rotational, inversional) changes or

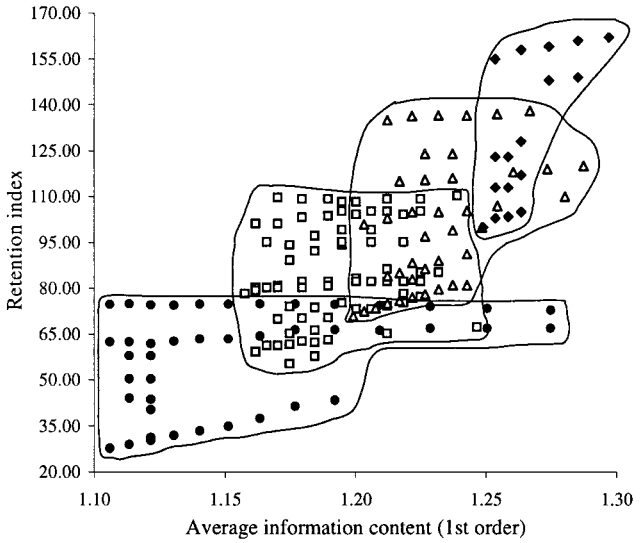

Figure 2. Retention index vs average information content. The distribution and overlap areas of the compounds. ●, monomethyl; □, di-methyl; △, tri-methyl; ♦, tetra-methyl.

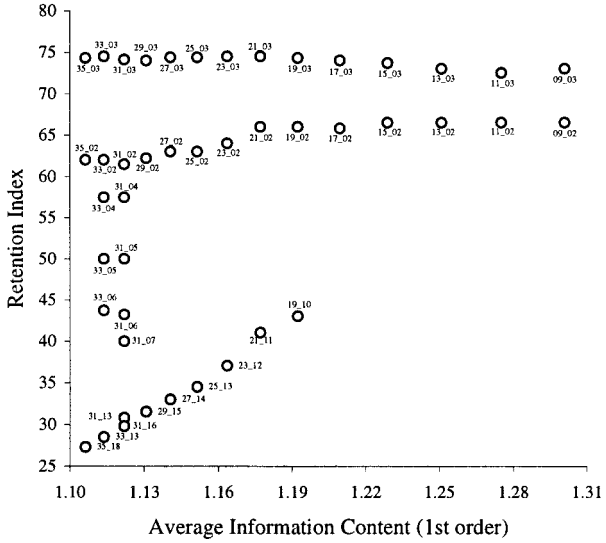

Figure 3. Dependence of retention index on the structure of the compound in monomethyl alkanes.

atomic reactivity in the molecule.[32] For the current set of data, this descriptor groups compounds into three sets: (i) 3, 3X, 3XY, 3XYZ substituted alkanes into one group, (ii) 4, 4X, 4XY, 4XYZ substituted alkanes into a second group, and (iii) all other compounds into a third group (in these designations X, Y, and Z refer to the positions of the second, third, and fourth methyl group,

(32) Strouf, O. *Chemical Pattern Recognition*; Wiley: New York, 1986.

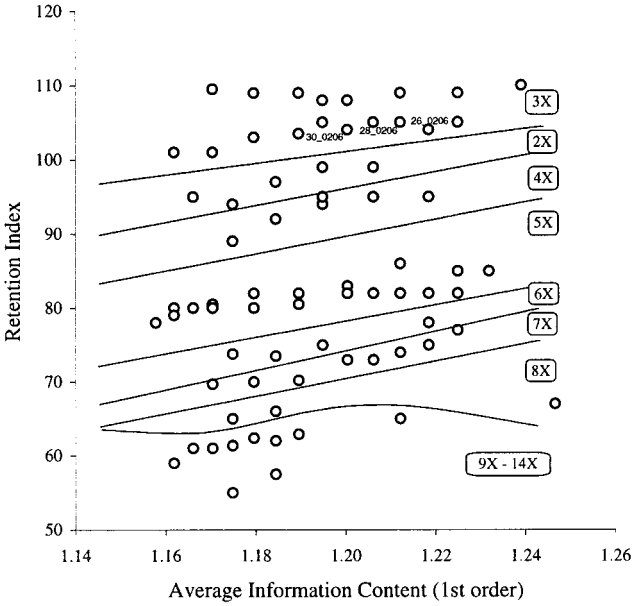

Figure 4. Dependence of retention index on the structure of the compound in dimethyl alkanes.

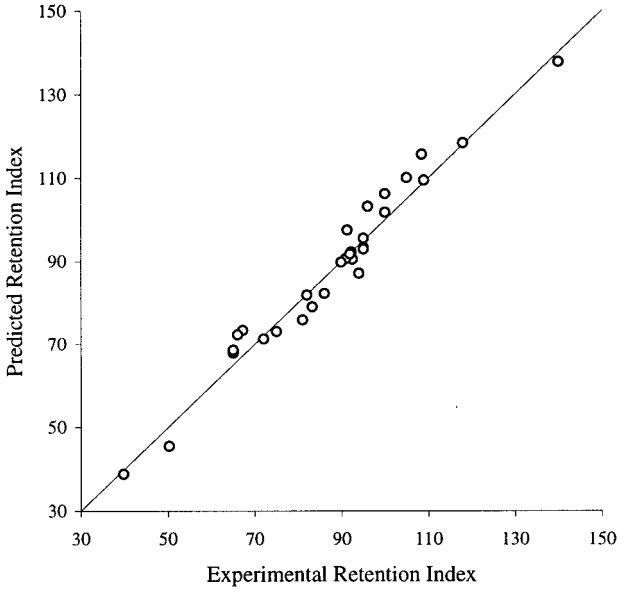

Figure 5. Plot of the predicted values for the external test set vs the experimental retention indexes.

respectively, in the backbone). Consequently this descriptor behaves as an indicator descriptor and shows different rotation and/or inversion behavior of 3- and 4-substituted compounds in comparison with other compounds.

$$J = \left(\frac{q}{\mu + 1}\right)\sum_{i,j}^{q}(S_i S_j)^{-1/2} \qquad (4)$$

The fourth descriptor, the Balaban index, is defined by eq 4,[33] where q is the number of edges in the molecular graph, n is the number of vertexes in the graph, $\mu = q - n + 1$ is the cyclometric number, and s_i, s_j are the distance sums (or distance degrees),

(33) Balaban, A. T. *Chem. Phys. Lett.* **1982**, *89*, 399−404.

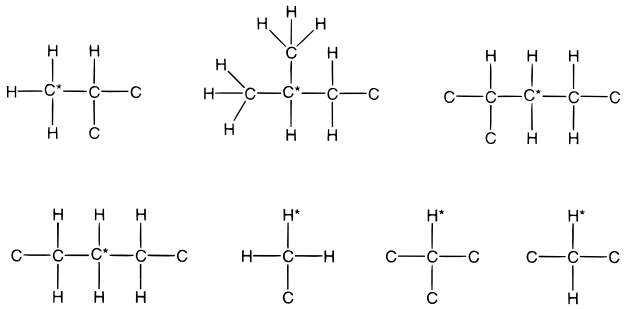

Figure 6. Seven additional connectivity patterns in the methyl-branched hydrocarbons in molecular graphs at coordination sphere two (the atom considered is marked with a star).

obtained by summation on the row i and column i (or row j and column j, respectively) of the distance matrix between atoms in the molecule. The descriptor reflects only the carbon backbone and, unlike the two information content indexes, neglects connected hydrogens. The Balaban Index is based on the molecular structure according to graph theory and the distance matrix and reflects the relative connectivity and effective size of the carbon chain to which are attached multiple methyl groups. For the current set of compounds this double feature distinguishes as four groups the mono-, di-, tri-, and tetrasubstituted alkanes, respectively. The magnitude of this descriptor increases with (i) increase in branching and (ii) increase in the number of atoms in the molecule. The retention index increases in the same order of the four groups of mono-, di-, tri-, and tetramethyl alkanes, respectively.

To understand more clearly how the retention index depends on the structure of a molecule, one can examine property vs descriptor relationships. Analyzing this relationship for ^{1}AIC (Figure 2) reveals some general trends. As already mentioned, the retention index depends (i) on the length of the carbon backbone and (ii) on the positions of the methyl groups connected to the backbone. The mono-, di-, tri-, and tetrasubstituted groups can be clearly distinguished but their retention indexes overlap as shown in Figure 2. Analysis of the regularities within each group clearly shows that the retention index also depends on the positions of methyl groups on the carbon backbone.

Within the group of the monomethyl alkanes (Figure 3), the 3-methylalkanes possess the highest retention indexes which are relatively constant for 3-methylalkanes with change of length of the carbon chain, varying only in the range of 72.5−74.5. 2-Methylalkanes have lower retention indexes than 3-methylalkanes, varying in the range 61.5−66.5. 4-, 5-, 6- and 7-Methylalkanes have still lower retention indexes. Retention indexes decrease in the order 3-, 2-, 4-, 5-, 6-, and 7-methyl-substituted alkanes. The fact that the retention indexes for 3-methylalkanes are not intermediate between the 2 and 4 analogues can be explained by the different conformation of the terminal ends of these compounds. When the position of the methyl group is close to the midsection of the carbon backbone, the retention indexes are the lowest (compounds 35_18_19_10). The following can be concluded from Figure 3: (i) the retention index depends on the position of the methyl group on the backbone; (ii) retention is lower when the substitution moves closer to the middle area in the backbone; (iii) 3-methylalkanes do not follow the general trend and show the highest retention index within a group; (iv) the

Table 6. Regularities in the Structure of Substituted Alkanes in the Range of Retention Units from 27 to 162

range up to	compounds			
	monomethyl	dimethyl	trimethyl	tetramethyl
40	23_12, 25_13, 27_14, 29_15, 31_07, 31_13, 31_16, 33_13, 33_17, 35_18			
50	19_10, 21_11, 31_05, 31_06, 33_05, 33_06			
60	31_04, 33_04	32_1418, 34_1317, 37_1323		
70	09_02, 11_02, 13_02, 15_02, 17_02, 19_02, 21_02, 23_02, 25_02, 27_02, 29_02, 31_02, 33_02, 35_02	22_0822, 27_0919, 31_1121, 32_0812, 32_0921, 33_0717, 33_1123, 34_0812, 34_1222, 35_0717, 35_0921, 36_1323		
80	09_03, 11_03, 13_03, 15_03, 17_03, 19_03, 21_03, 23_03, 25_03, 27_03, 29_03, 31_03, 33_03, 35_03	25_0711, 26_0610, 26_0711, 27_0713, 28_0713, 29_0717, 30_0610, 31_0711, 32_0610, 33_0517, 34_0610, 35_0509, 36_0517, 37_0509, 37_0517, 38_0517	33_111519, 34_121620, 35_131721, 36_141822, 37_151923, 38_162024, 39_151923, 40_141822	
90		24_0509, 25_0511, 25_0517, 26_0511, 27_0511, 27_0517, 28_0515, 29_0513, 29_0519, 31_0513, 31_0517, 33_0519, 34_0416, 35_0519	31_111519, 32_121620, 33_071115, 34_081216, 35_071115, 35_131723, 36_081216, 37_071319	
100		26_0408, 28_0210, 28_0410, 30_0210, 30_0212, 30_0410, 32_0208, 32_0408, 34_0210, 36_0212	30_061418, 31_071317, 32_061418, 34_061418	38_10141822
110		23_0309, 25_0311, 25_0315, 26_0206, 27_0307, 27_0315, 28_0206, 29_0307, 29_0313, 30_0206, 30_0307, 31_0307, 31_0313, 31_0315, 33_0309, 33_0315, 35_0307, 35_0315, 37_0315	25_050913, 29_051317, 31_051317, 33_051317, 35_050913, 37_051317, 39_051317	35_11151924, 36_10141822, 37_11151924
120			24_040812, 26_040812, 28_040812, 32_041216, 34_040812, 36_040816	35_09131721, 36_08121620, 37_09131721
130			32_021016, 34_021016	35_07111519, 36_06101216, 37_07111519
140			27_030711, 29_030711, 31_030711, 33_030715, 35_030715, 37_030715	
150				31_04081216, 33_04081216
160				37_03071115, 35_03071115, 33_03071115
162				29_03071115, 31_03071115

retention index for the methyl group at position 10 and higher on long chains of 35−19 carbons in the backbone increases from 27.3 to 43.

Dimethylalkanes (2X to 14X) follow a pattern similar to that for the monomethylalkanes. Those with one methyl group at the 3-position possess the highest retention indexes in the range of 101−110. 2,X-Dimethyl compounds have retention indexes in the range of 94−97. The only exceptions are 2,6-dimethylalkanes, with indexes overlapping with those for the 3,X-compounds, in the range 104−105. This can be explained by the conformation near the end of the chain of these compounds, similar to those of 3-methyl-substituted alkanes. 4,X-Dimethyl compounds have indexes in the range 89−95, near to that for the 2,X-compounds. The least regular are the 5,X-dimethyl compounds in the range 78−86. Compounds of the 6,X-, 7,X-, and 8,X-classes show retention indexes in ranges of 73.5−78, 69.7−77, and 65−67, respectively. The 9,X- to 14,X-dimethyl compounds possess still lower retention indexes in the range 55−65. Tri- and tetrasubstituted compounds follow rules similar to those of the previously discussed mono- and trimethyl analogues.

To test the quality of our correlation equation, the retention indexes for 30 methylalkanes not used for building the QSPR model were predicted. The compounds in the external test set (Table 3) were measured by using the same methodology as described above. Then the appropriate descriptor values were inserted into the correlation equation (Table 4), and the respective retention indexes were calculated.[1] The predicted retention indexes are shown in Table 3 and plotted against the experimental values in Figure 5, with $R^2 = 0.962$. The average error of the whole set of 178 compounds is 4.6; however, if we exclude the 13 monomethyl compounds of less than 20 carbon atoms, we get for the remaining 165 compounds an average prediction error of 4.3. The average error for the external set is 3.0 (standard deviation is 4.0) and is in line with the average experimental error

of 4.0. These results clearly demonstrate the practical utility of the study undertaken.

The prediction of the structures which correspond to a particular retention index is more challenging and is complicated by the overlapping of structures. However, some estimates can be given on the basis of the arrangement of the structures in Figure 1 as given in Table 6. For a retention index of up to 50, only monomethylalkanes are expected and for retention indexes of between 130 and 140 only 3,X,Y-trimethylalkanes. Retention indexes of above 140 imply tetramethylalkanes. Structure diversity for retention indexes in the range 50−130 is much greater. In the range from 50 to 60, we expect mainly 4-monomethylalkanes and dimethylalkanes with one substituent close to the midsection of the backbone. Retention indexes of 60−70 include 2-monomethylalkanes and dimethylalkanes similar to those in the previous range. Retention indexes from 70 to 80 cover three classes (i) 3-monomethylalkanes, (ii) 5,X-, 6,X-, and 7,X-dimethylalkanes, and (iii) trimethylalkanes where the first substituent is close to the midsection of the backbone. The range from 80 to 90 includes (i) 5,X-dimethylalkanes with one additional compound (Table 6) and (ii) trimethylalkanes. The next range, 90−100, also covers three classes: (i) 2,X- and 4,X-dimethylalkanes, (ii) 6,X,Y-trimethylalkanes, and (iii) tetramethylalkanes. The range 100−110 includes 3,X-dimethylalkanes (with the exception of 2,6-dimethylalkanes); 5,X,Y-trimethylalkanes overlap with the previous region. The next two intervals include di- and trimethylalkanes. In the range of 110−120 one can expect 4,X,Y-trimethylalkanes along with 8,X,Y,Z- and 9,X,Y,Z-trimethylalkanes. In the range of 120−130, 2,X,Y-trimethylalkanes along with 6,X,Y,Z- and 7,X,Y,Z-trimethylalkanes appear.

It was to be expected that the GC retention index in methylalkanes should be modeled by molecular structural descriptors that reflect the relative position and the number of the methyl groups attached to the carbon backbone, the conformation of the compound, and the length of the carbon backbone. As our QSPR model shows, these structural differences are best described by the molecular graphs utilized in topological descriptors and supported by quantum-chemical descriptors. In other words, topological factors govern the chromatographic retention behavior of methylalkanes. Our study also shows that even small differences in structure can significantly change physical and chemical properties.

Intermolecular solute−solute and solute−stationary phase interactions are known to play an important role in determining the GC retention index. The electronic environment of the molecules involved is usually also important, but for the current set of methylalkanes and a nonpolar stationary phase, electronic effects are not significant. On the other hand, the solute−solute interactions that depend on conformation of the structures may be important and can probably be related to the formation of a liquid monolayer while gas is eluted through the column. Currently this remains as a working hypothesis for further studies, which should seek additional insight into descriptors, already available or yet to be developed, that can describe such behavior and the possibility of successfully incorporating them into improved QSPR models.

CONCLUSION

A quantitative structure−property relationship model was derived to study the GC retention index of methyl-branched alkanes for a diverse set of 178 compounds produced by insects. A four-descriptor equation was developed with a squared correlation coefficient of 0.9585 and a standard error of 5.8, which is close to the average experimental error of 4. Topological descriptors are found to have high coding capabilities for the GC retention index and are selected to represent the chemical structures effectively and simply. The correlation equation and descriptors can be used for the prediction of the retention index for unknown structures. The treatment also classifies structures and discloses general trends for which types of structure can be expected if only the retention index is known.

ACKNOWLEDGMENT

We thank Dr. Yilin Wang for her contribution to this work. This research was supported in part by a grant from NATO (GRC:SA.5.2.25).

SUPPORTING INFORMATION AVAILABLE

The intercorrelations between the four descriptors and a sample to calculate and predict the KI value of 29_0509 are given. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review July 20, 1999. Accepted October 13, 1999.

AC990800W